Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 64102, 20 pages
doi:10.1155/2007/64102

Research Article

A Comprehensive Noise Robust Speech Parameterization Algorithm Using Wavelet Packet Decomposition-Based Denoising and Speech Feature Representation Techniques

Bojan Kotnik and Zdravko Kačič

Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova ul. 17, 2000 Maribor, Slovenia

Received 22 May 2006; Revised 12 January 2007; Accepted 11 April 2007

Recommended by Matti Karjalainen

This paper concerns the problem of automatic speech recognition in noise-intense and adverse environments. The main goal of the proposed work is the definition, implementation, and evaluation of a novel noise robust speech signal parameterization algorithm. The proposed procedure is based on time-frequency speech signal representation using wavelet packet decomposition. A new modified soft thresholding algorithm based on time-frequency adaptive threshold determination was developed to efficiently reduce the level of additive noise in the input noisy speech signal. A two-stage Gaussian mixture model (GMM)-based classifier was developed to perform speech/nonspeech as well as voiced/unvoiced classification. The adaptive topology of the wavelet packet decomposition tree based on voiced/unvoiced detection was introduced to separately analyze voiced and unvoiced segments of the speech signal. The main feature vector consists of a combination of log-root compressed wavelet packet parameters, and autoregressive parameters. The final output feature vector is produced using a two-staged feature vector postprocessing procedure. In the experimental framework, the noisy speech databases Aurora 2 and Aurora 3 were applied together with corresponding standardized acoustical model training/testing procedures. The automatic speech recognition performance achieved using the proposed noise robust speech parameterization procedure was compared to the standardized mel-frequency cepstral coefficient (MFCC) feature extraction procedures ETSI ES 201 108 and ETSI ES 202 050.

Copyright © 2007 B. Kotnik and Z. Kačič. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Automatic speech recognition (ASR) systems have become indispensable integral parts of modern multimodal man-machine communication dialog applications such as voice-driven service portals, speech interfaces in automotive navigational and guidance systems, or speech-driven applications in modern offices [1]. As automatic speech recognition systems are evolutionally moving from controlled laboratory environments to more acoustically dynamic places, noise robustness criteria must be assured in order to maintain speech recognition accuracy above a sufficient level. If a recognition system is to be used in noisy environments it must be robust to many different types and levels of noise, categorized as either additive/convolutive noises, or changes in the speaker's voice due to environmental noise (Lombard's effect) [1, 2]. Two large groups of noise robust techniques are commonly used in modern automatic speech recognition systems. The first one comprises noise robust speech parameterization techniques and the second group consists of acoustical model compensation approaches. In both cases, the methods for robust speech recognition are focused on minimization of the acoustical mismatch between training and testing (recognition) environments. Namely, this mismatch is the main reason for the degradation of automatic speech recognition performance [1, 3, 4]. This paper focuses on the first group of noise robust techniques: on noise robust speech parameterization procedures. Development of the following algorithms needs to be considered with the aim of improving automatic speech recognition performance under adverse conditions: (1) compact and reliable representation of speech signals in the time-frequency plane, (2) efficient signal-to-noise ratio (SNR) enhancement or denoising algorithms to cope with various colored and nonstationary additive noises as well as channel distortion (convolutional noises), (3) accurate voice activity detection strategies are necessary to implement a frame-dropping principle and to discard noise-only frames, (4) effective feature postprocessing algorithms should be applied to transform feature vectors to the lower-dimensional space, to decorrelate elements in feature vectors, and to enhance the accuracy of the classification process.

This article presents a novel noise robust speech parameterization algorithm, shortly denoted as WPDAM, using joint wavelet packet decomposition and autoregressive modeling. The proposed noise robust front-end procedure produces solutions for all the four noise robust speech parameterization issues mentioned above and should, therefore, achieve better automatic speech recognition performance in comparison with the standardized mel-frequency cepstral coefficient (MFCC) feature extraction procedure [5, 6].

MFCCs [7], derived on the basis of short time Fourier transform (STFT) and power spectrum estimation, have been used to date as fundamental speech features in almost every state-of-the-art speech recognition system. Nevertheless, many authors have reported on the drawbacks of the MFCC speech parameterization technique [1, 8–12]. The windowed STFT was one of the first transforms to provide temporal information about the frequency content of signals [13, 14]. The STFT-based approach has, due to constant analysis window length (typically 20–32 milliseconds), fixed time-frequency resolution and is, therefore, not optimized to simultaneously analyze the nonstationary and quasi-stationary parts of a speech signal with the same accurateness [15–18].

Speech is a highly dynamic process. A multiresolutional approach is needed in order to achieve reliable representation of the speech signal in the time-frequency plane. Instead of using fixed-resolution STFT, a wavelet transform can be used to efficiently represent the speech signal in the time-frequency plane [17, 18]. Wavelet transform (WT) has become a popular tool in many research domains. It decomposes data into a sparse, multiscale representation. The wavelet transform, with its flexible time-frequency resolution is, therefore, an appropriate tool for the analysis of signals having both short high-frequency bursts and long quasi-stationary components [19].

Examples of WT usage in the feature extraction process can be found in [8, 10, 20]. The wavelet packet decomposition tree (WPD), which tries to mimic the filters arranged in the Mel scale, in a similar fashion to that achieved by the MFCC has already been used in [21]. It is shown that the usage of WPD prior to the feature extraction stage leads to a performance improvement in the automatic speaker identification system [9, 21] or automatic speech recognition system when compared to the baseline MFCC system [9]. Optimal structure for the WPD tree using an entropy based measure has been proposed [15, 22] in the research area of signal coding. It has been shown that entropy based optimal coding provides compact coding of the signals, while losing a minimum of the useful information [23].

Different denoising strategies based on speech signal representation using wavelets can be found in literature [18, 19, 21, 24–27]. One of the objectives of the proposed noise robust speech parameterization procedure is also the development of a computationally efficient improved alternative—a denoising algorithm based on modified soft thresholding strategy with the application of time-frequency adaptive threshold and adaptive thresholding strength.

The rest of this article is organized as follows: Sections 2–9 provide, together with its subsections, a detailed description of all processing steps applied in the proposed noise robust feature extraction algorithm WPDAM. The automatic speech recognition performance of the proposed algorithm is evaluated using Aurora 2 [28–30] and Aurora 3 [31–34] databases and compared to the ETSI ES 201 108 and ETSI ES 202 050 standard feature extraction algorithms [5, 30, 35]. Section 10 gives a description of the performed experiments, corresponding results and discussions. The performance comparison to other complex front ends, as well as the computational requirements will also be provided. Finally, Section 11 concludes the paper.

2. DEFINITION OF PROPOSED ALGORITHM WPDAM

The block diagram for the proposed noise robust speech parameterization procedure is presented in Figure 1. In the first step, the digitized input speech signal is segmented into overlapping frames, each of length 48 milliseconds with a frame shift interval of 10 milliseconds. The overlapping frames represent the basic processing units of all the processing steps in the proposed algorithm. In the second step, a speech preprocessing procedure is applied. It consists of high-pass filtering with a cutoff frequency of 70 Hz. Afterwards, a speech pre-emphasis is applied. It boosts the higher frequency contents of the speech signal and, therefore, improves the detection and representation of the low-energy unvoiced segments of the speech signal, which dominate mainly in the high-frequency regions. The third processing step applies a wavelet packet decomposition of the preprocessed input signal. Wavelet packet decomposition (WPD) is used to represent the speech signal in the time-frequency plane [17, 18]. In the next stage, a voice activity and voiced-unvoiced detections are applied, preceded by a preliminary additive noise reduction scheme using time-frequency adaptive threshold and smoothed modified soft thresholding procedure. After preliminary denoising, the denoised speech signal is reconstructed. Then the autoregressive parameters of the enhanced speech signal are extracted and linear prediction cepstral coefficients (LPCC) are computed. The feature vector constructed on the basis of LPCCs is applied in the statistical classifier used in the voice activity detection procedure. This classifier is based on Gaussian mixture model (GMM). In the training phase, the GMM models for "speech" and "non-speech" are trained and later, in the test phase, these two models are evaluated using the feature vector of a particular frame of the input speech signal. The emission probabilities of the two GMM models are smoothed in time and compared. The classification result is binary and defined with that particular GMM model, which generates the highest emission probability. The voiced-unvoiced detection, which is performed for speech-only frames, uses the same principle of

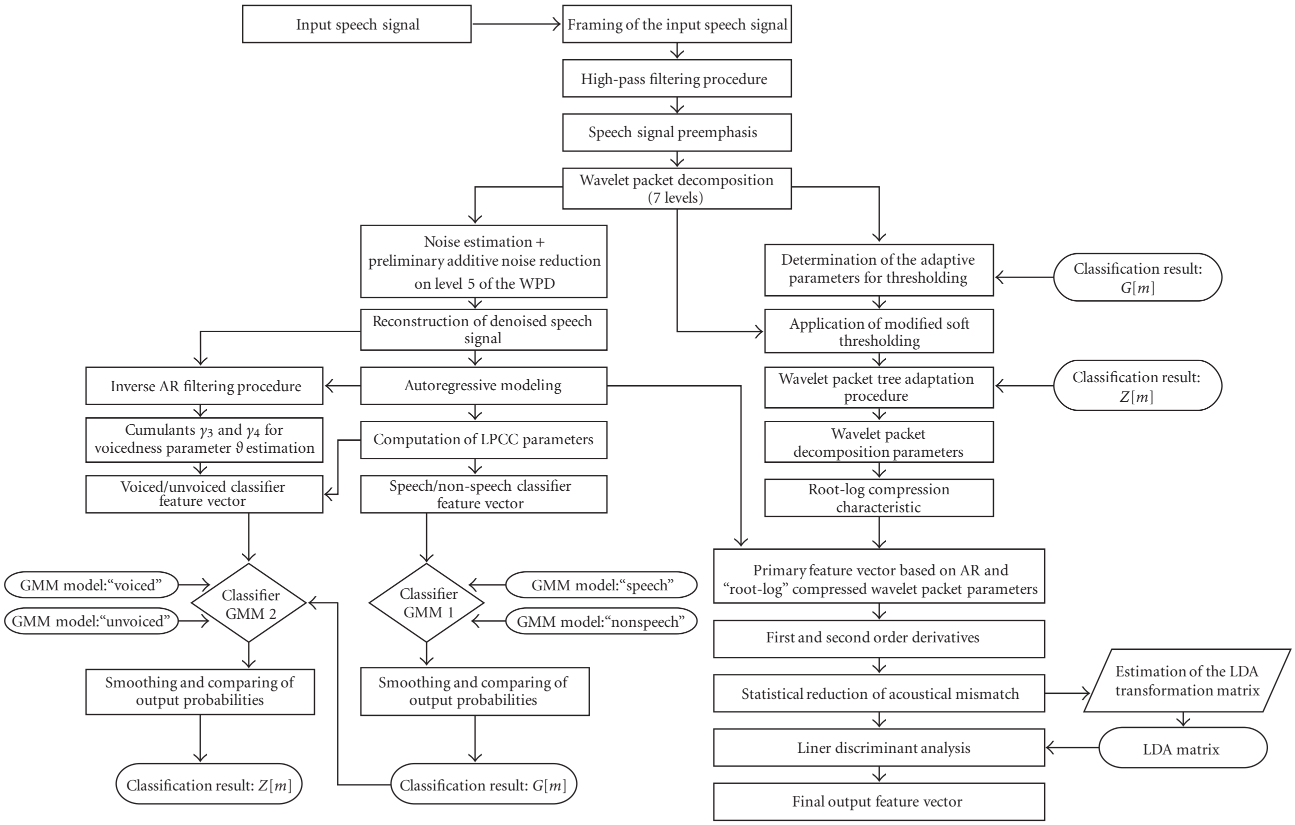
Input speech signal
Framing of the input speech signal
High-pass filtering procedure
Speech signal preemphasis
Wavelet packet decomposition (7 levels)
Noise estimation + preliminary additive noise reduction on level 5 of the WPD
Reconstruction of denoised speech signal
Inverse AR filtering procedure
Autoregressive modeling
Cumulants γ_3 and γ_4 for voicedness parameter ϑ estimation
Computation of LPCC parameters
Voiced/unvoiced classifier feature vector
Speech/non-speech classifier feature vector
GMM model:"voiced"
GMM model:"unvoiced"
Classifier GMM 2
Classifier GMM 1
GMM model:"speech"
GMM model:"nonspeech"
Smoothing and comparing of output probabilities
Smoothing and comparing of output probabilities
Classification result: $Z[m]$
Classification result: $G[m]$
Determination of the adaptive parameters for thresholding
Classification result: $G[m]$
Application of modified soft thresholding
Wavelet packet tree adaptation procedure
Classification result: $Z[m]$
Wavelet packet decomposition parameters
Root-log compression characteristic
Primary feature vector based on AR and "root-log" compressed wavelet packet parameters
First and second order derivatives
Statistical reduction of acoustical mismatch
Estimation of the LDA transformation matrix
Liner discriminant analysis
LDA matrix
Final output feature vector

Figure 1: Block diagram of proposed noise robust speech parameterization algorithm WPDAM.

statistical classification. The only difference is a modification of the input feature vector, which is constructed from autoregressive parameters with an added special voiced/unvoiced feature. The voicing feature is represented by the ratio of the higher-order cumulants of the LPC residual signal. The main wavelet-based denoising procedure uses a more advanced time-frequency adaptive threshold determination procedure. The speech/nonspeech decision and principles of minimum statistics are also used. Once the threshold is determined, the thresholding process is performed. The two modified soft thresholding characteristics are introduced: piecewise linear modified soft thresholding (preliminary denoising) and smoothed modified soft thresholding characteristic (primary speech signal denoising).

The primary features are represented by the wavelet packet decomposition parameters of the denoised input speech signal. The parameters are estimated on the basis of the wavelet packet decomposition tree's adaptive topology, using voiced-unvoiced decision. The wavelet packet parameters are compressed using the proposed combined root-log compression characteristics. The primary feature vector consists of a combination of compressed wavelet packet parameters, and of autoregressive parameters. The global frame energy of the denoised input speech signal is also added, as the last element of the primary feature vector. Next, the dynamic features—the first- and second-order derivatives of the statistical elements—are also added to the final feature vector. The first step in the feature vector postprocessing consists of a procedure for the statistical reduction of the acoustical mismatch between the training and testing conditions. The final output feature vector is computed using linear discriminant analysis (LDA).

The proposed noise-robust feature extraction procedure consists of training and testing phases. In the training phase, the statistical GMM models (speech/nonspeech and voiced/unvoiced GMMs), the parameters for statistical mismatch reduction, and LDA transformation matrix need to be estimated before the actual usage of the proposed algorithm in the feature extraction process.

3. INPUT SPEECH SIGNAL PREPROCESSING PROCEDURE

The main purpose of speech signal preprocessing is the elimination of primary disturbances in the input signal, as well as optimal preparation of the speech signal for further processing steps, with the aim of achieving higher automatic speech recognition accuracy. The proposed preprocessing procedure consists of high-pass filtering, and pre-emphasis of the input speech signal. A high-pass filter with a cut-off frequency f_c of around 70 Hz is proposed with the aim of eliminating the unwanted effects of low-frequency disturbances. Namely, the speech signal does not contain useful information in the frequency band from 0 to 70 Hz and, therefore, the frequency content in that band can be strongly attenuated. A Chebbyshev infinite impulse response (IIR) filter of type 1 was constructed in order to achieve a fast transit from the stop to passband of the proposed low-order highpass filter. The proposed filter has a passband ripple of, at most, 0.01 dB.

The perceptual loudness of the human auditory system depends on the frequency contents of the input sound wave. It is commonly known that the unvoiced sounds contain less energy than the voiced segments of speech signals [2]. However, the correct and accurate detection and classification of unvoiced phonemes is also of crucial importance when achieving the highest automatic speech recognition results [1, 20]. Therefore, speech pre-emphasis techniques were introduced to improve the acoustic modeling and classification process of the unvoiced speech signal segments [13, 14]. The MFCC standardized feature extraction procedure ETSI ES 201 108 [5] uses the first-order pre-emphasis filter, as described in the transfer function $H_P(z) = 1 - \alpha z^{-1}$. A new pre-emphasis filter $H_{\text{PREEMPH}}(z)$ is proposed for the presented WPDAM. The proposed pre-emphasis filter does not modify the frequency content of the input signal in the frequency region from 0 to 1 kHz. For the frequencies from 1 kHz up to 4 kHz (the sampling frequency of $f_S = 8$ kHz is presumed) the amplification of the input speech signal is progressively increased and achieves its maximum at 3.52 dB, at a frequency of 4 kHz.

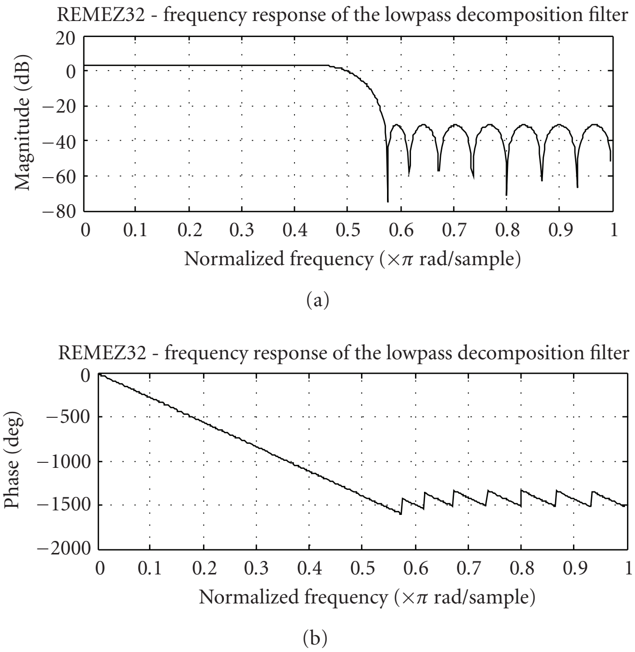

Figure 2: Frequency response of the REMEZ32.

4. WPD-BASED SPEECH SIGNAL DENOISING PROCEDURE

The environmental noises surrounding the user of the voice-driven applications represent the main obstacle to achieve a higher degree of automatic speech recognition accuracy [1, 24, 36–39]. Modern automatic speech recognition systems are based on a statistical approach using hidden Markov models and, therefore, their efficiency depends on the degree of acoustical match between training and testing environments [1, 14]. If the training of acoustical models is performed using studio-quality speech with the highest SNR, and if, in practical usage, the input speech signal is captured in a low SNR environment (interior of driven car on the highway, e.g.), then a significant degradation of the speech recognition performance is to be expected. However, it should be noted that increased SNR does not lead always to the improvements in the ASR performance. Therefore, the main goal of presented additive noise reduction principles is the reduction of acoustic mismatch between the training and testing environments [1].

4.1. Definition of the WPD applied in the proposed denoising procedure

Discrete-time implementation of the wavelet transform is defined as the iteration of the two-channel filterbank, followed by a decimation-by-two unit [16–18]. Unlike the discrete wavelet transform (DWT), which is obtained by iterating on the lowpass branch only, the filterbank tree can be iterated on either branch at any level, resulting in a tree-structured filterbank called a wavelet packet filterbank tree [18]. In the proposed noise robust feature extraction WPDAM, a J-level WPD algorithm is applied to decompose the high-pass filtered and pre-emphasized signal $y[n, m]$, where n and m are the sample and the frame indexes, respectively.

The nomenclature used in the presented article is as follows: the WPD level index is denoted by j whereas the wavelet packet (subband) index is represented by k. The wavelet packet sequence of frame m on level j and subband k is represented by $W_{j,k}^{m}$. The decomposition tree consists of J decomposition levels and has a total of N_{NODE} nodes. K output nodes exist, where $K = 2^J$. The wavelet function REMEZ32 is applied in the presented feature extraction algorithm WPDAM. The REMEZ32 is based on equiripple FIR filter definition performed using the Parks-McClellan optimum filter design procedure with Remez's exchange algorithm [40, 41]. The impulse response length of the proposed filter is equal to the length of classical wavelet function Daubechies-16 (32 taps) [16]. Figures 2 and 3 present the frequency response and corresponding wavelet function of the REMEZ32, respectively. Note that the mother wavelet function presented on Figure 3 is based on 3-times interpolated impulse response of the high-pass reconstruction filter REMEZ32 (hence the length of 96 taps on Figure 3). The filter corresponding to REMEZ32 has linear phase response and magnitude ripples of constant height. The transition band of the magnitude response is much narrower (280 Hz) than the transition band at Daubechies-16 (1800 Hz), but the final attenuation in the stop band (−32 dB) is smaller than that at the Daubechies-16 (−300 dB) [16, 41].

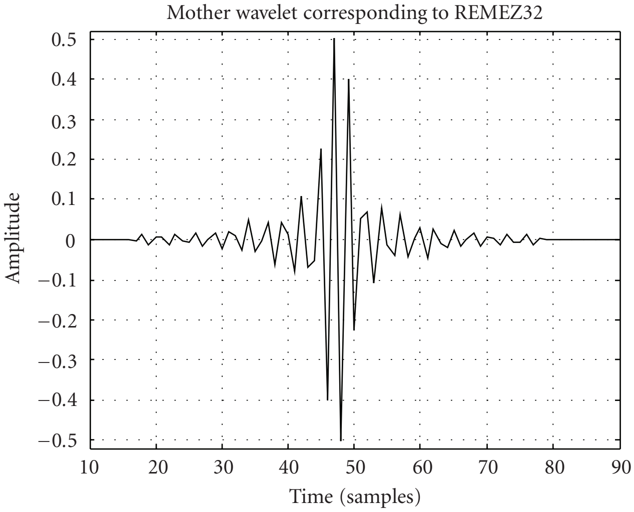

Figure 3: Wavelet function REMEZ32.

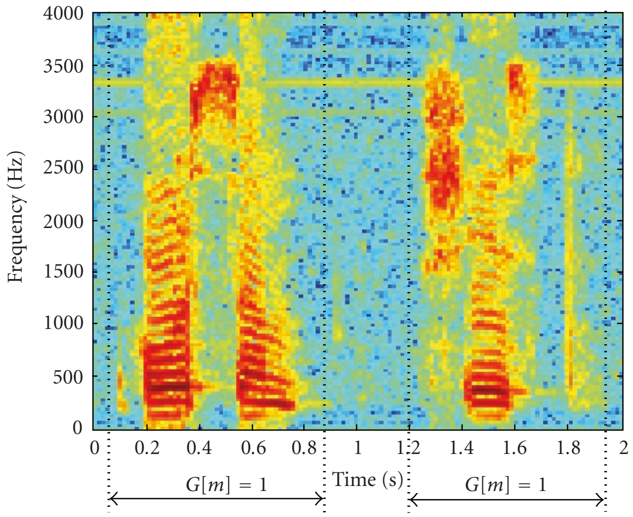

Figure 4: Time-frequency representation of speech signal with denoted voice activity detection borders.

4.2. *The definition of proposed time-frequency adaptive threshold*

The main goal of the proposed WPD-based noise reduction scheme is achievement of the strongest possible signal-to-noise ratio (SNR) improvement at lowest additional signal distortion [21, 25, 27, 36, 42]. The compromising solution is achievable only with accurate time-frequency adaptive threshold estimation procedure, and with definition of efficient thresholding algorithm.

Figure 4 shows a speech signal spectrogram with added voice activity decision borders. It is evident from this spectrogram that, even in the speech region ($G[m] = 1$), not all of the frequency regions contain useful speech information. Therefore, it can be speculated that the noise spectrum can be effectively estimated not only in the pure-noise regions ($G[m] = 0$) but also inside the speech regions ($G[m] = 1$). The main principles of this minimum statistics approach [38] will be used in the development of the proposed threshold determination procedure. The presented noise reduction procedure only operates on output nodes of the lowest level of the wavelet packet decomposition tree, which is defined here by $j = 7$. The adaptive threshold $T_k^j[m]$ determination method is performed as follows. For each frame m of the input speech signal $y[m,n]$, the Donoho's [25] threshold $\mathrm{DT}_k^j[m]$ is computed at every output k of the lowest wavelet packet decomposition level j:

$$\begin{aligned} \mathrm{DT}_k^j[m] &= \sigma_k^j[m]\sqrt{2\log\left(N_k^j\right)},\\ \text{where } \sigma_k^j[m] &= \frac{1}{\gamma_{\mathrm{MAD}}}\ \mathrm{Median}\left(\left|W_k^j(x[m,n])\right|\right). \end{aligned} \tag{1}$$

When the SNR of the input noisy speech signal $y[n]$ is relatively low (SNR < 5 dB), high inter-frame fluctuations in the threshold value result in additional distortion of the denoised speech signal, which are similar to musical noises—artefacts known in spectral subtraction algorithms [19, 36, 38]. These abrupt changes in inter-frame threshold values can be reduced using the following first-order autoregressive-smoothing scheme:

$$\overline{\mathrm{DT}_k^j[m]} = (1-\delta)\,\mathrm{DT}_k^j[m] + \delta\overline{\mathrm{DT}_k^j[m-1]}, \tag{2}$$

where the smoothing factor δ has a typical value from the interval $(0.9, 1.0]$. The final time-frequency adaptive threshold $T_k^j[m]$ is produced using the smoothed Donoho's threshold $\overline{\mathrm{DT}_k^j[m]}$, and voice activity decision $G[m]$ as follows.

(i) If the current frame m does not contain useful speech information ($G[m] = 0$), then the proposed time-frequency adaptive threshold $T_k^j[m]$ is equivalent to the value of the smoothed Donoho's threshold

$$T_k^j[m] = \overline{\mathrm{DT}_k^j[m]}, \quad \text{if } G[m] = 0. \tag{3}$$

(ii) If the current frame m corresponds to the speech segment S of the input signal ($G[m] = 1$ and $m \in S$), then the threshold $T_k^j[m]$ is determined using the minimum-statistic principle: inside the speech segment S, the interval I of the length of D frames is selected, where $I = [m - D/2, m + D/2]$, and $I \subseteq S$. For the frame m, wavelet packet decomposition level j, and node k, the threshold $T_k^j[m]$ corresponds to the minimal smoothed Donoho's threshold value $\overline{\mathrm{DT}_k^j[m']}$, where m' runs over all values from the interval I:

$$\begin{aligned} T_k^j[m] &= \operatorname*{Min}_{m'\in I}\left(\overline{\mathrm{DT}_k^j[m']}\right),\\ \text{where } I &= \left[m - \frac{D}{2}, m + \frac{D}{2}\right], \quad I \subseteq S. \end{aligned} \tag{4}$$

The proposed time-frequency adaptive threshold $T_k^j[m]$ is used, together with the proposed modified soft thresholding algorithm (presented in the following subsection), to reduce the level of additive noise in the input noisy speech signal $y[n,m]$.

4.3. Modified soft thresholding algorithm

The selection of the thresholding characteristics has strong impact on the quality of the denoised output speech signal [25, 27]. Detailed analysis of well-known hard and soft thresholding techniques showed that there are two main reasons why the distortion of the denoised output speech signal occurs [21]. The first reason is the strong discontinuity of the input-output thresholding characteristics, and the second reason is setting to zero those coefficients, the absolute values of which are below the threshold. Most of the speech signal's energy is concentrated at lower frequencies (voiced sounds), whereas the unvoiced low-energy segments of the speech signal are mainly located at higher frequencies [2, 43]. The wavelet coefficients of the unvoiced speech are, due to its lower amplitude, more masked by surrounding noise and, therefore, they are easily attenuated by inappropriate thresholding operations such as hard or even soft thresholding [27]. In the proposed smoothed modified soft thresholding technique, special attention is dedicated to unvoiced regions inside the speech signal and, therefore, those wavelet coefficients, the absolute values of which lie below the threshold value, are treated with special care. The proposed smoothed modified soft thresholding function has a smooth, nonlinear attenuating shape for the wavelet packet coefficients, the absolute values of which lie below the threshold. The smoothed modified soft thresholding function is defined by the following equation:

$$\begin{aligned}
&\text{IF } |W(x[n])| > T_k^j, \quad \text{THEN}\\
&\quad W(s'[n]) = W(x[n])\\
&\text{ELSE}\\
&\quad W(s'[n]) = T_k^j\left[\operatorname{sign}(W(x[n]))\frac{1}{\rho_k^j}\left((1+\rho_k^j)^{|W(x[n])|/T_k^j}-1\right)\right].
\end{aligned} \tag{5}$$

For greater readability, the frame index m was discarded from the equation above. The adaptive parameter $\rho_k^j[m]$ in (5) defines the shape of the attenuation characteristic for the wavelet packet coefficients, the absolute values of which lie below the threshold $T_k^j[m]$. The adaptive parameter $\rho_k^j[m]$ is determined as follows:

$$\rho_k^j[m] = \theta\frac{\max(|W_k^j(x[m,n])|)}{T_k^j[m]}. \tag{6}$$

The global constant θ is estimated on the basis of an analysis of the minimum mean square error (MMSE) $e[n]$ between the clean speech signal $s[n]$ and the estimated clean speech signal $s'[n]$: $e[n] = s[n] - s'[n]$. The clean speech signal must be known in order to estimate the parameter θ. Therefore, the speech database Aurora 2 [29] was applied in $\rho_k^j[m]$ estimation procedure, where the time-aligned clean and noisy signals of the same utterance are available. As evident from (6), the attenuation factor $\rho_k^j[m]$ depends on the threshold value $T_k^j[m]$, as well as on the maximum absolute value of the wavelet coefficient found in the wavelet packet coefficient sequence $W_k^j(x[m,n])$. By applying the presented smoothed modified soft thresholding operation, better quality of output denoised speech is expected especially in unvoiced regions, as in the cases of classical hard and soft thresholding techniques. The illustrative diagram in Figure 5 represents the two smoothed modified soft thresholding characteristics at two different values for adaptive parameter $\rho_k^j[m]$: $\rho_k^j[m] = 30$ and $\rho_k^j[m] = 600$. At lower values for the parameter $\rho_k^j[m]$, the attenuation of wavelet coefficients becomes less aggressive and, therefore, those wavelet coefficients with absolute values below the threshold are better preserved. Therefore, the information contained in lower-valued coefficients (probably in unvoiced regions) is retained better. In order to make the following steps possible, a partial reconstruction of the denoised signal is needed. Namely, in Section 6 the adaptive topology of the wavelet packet decomposition tree will be utilized. Therefore, the denoised speech signal up to the level $j = 4$ has to be reconstructed using already mentioned REMEZ32 reconstruction filter.

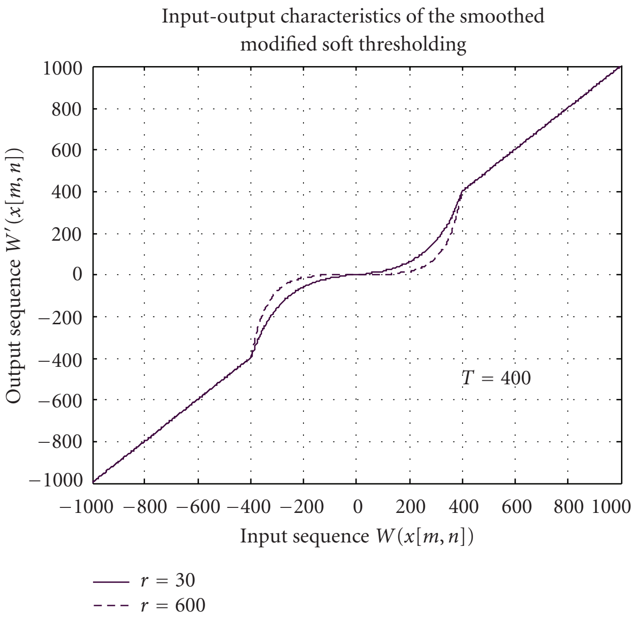

Figure 5: Two smoothed modified soft thresholding transfer characteristics.

5. SPEECH ACTIVITY AND VOICED/UNVOICED DETECTION

The main properties, which are demanded for voice activity and voicing detection (VAD) are reliability, noise robustness, accuracy, adaptation to changing operating conditions,

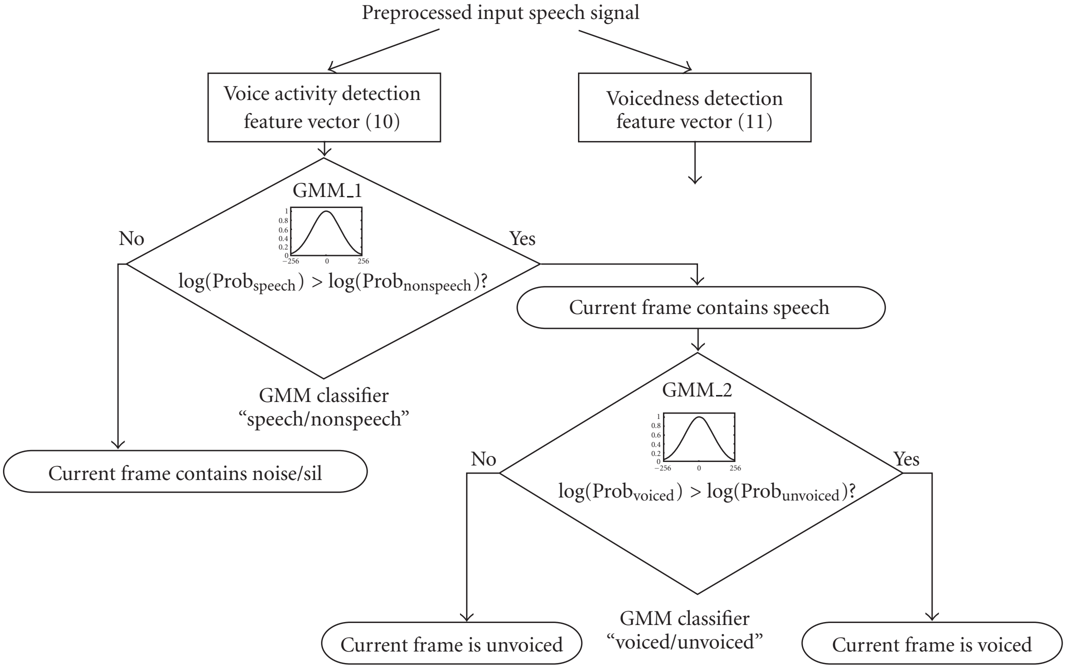
Preprocessed input speech signal
Voice activity detection feature vector (10)
Voicedness detection feature vector (11)
GMM_1
No
Yes
log(Prob$_{\text{speech}}$) > log(Prob$_{\text{nonspeech}}$)?
GMM classifier "speech/nonspeech"
Current frame contains speech
Current frame contains noise/sil
GMM_2
No
Yes
log(Prob$_{\text{voiced}}$) > log(Prob$_{\text{unvoiced}}$)?
GMM classifier "voiced/unvoiced"
Current frame is unvoiced
Current frame is voiced

Figure 6: Two-stage GMM-based statistical classification procedure.

speaker and speaking style independence, low computational and memory requirements, high operating speed (at least real-time operation), and reliable operation without a-priori knowledge about the environmental-noise characteristics [1, 28, 44–46]. The most problematic requirement of the VAD algorithm is robustness to different noises, SNRs, and adaptation of the VAD parameters to changing environmental characteristics [1, 44, 47]. The computationally most efficient VAD algorithms are based on signal energy estimation principles, zero crossing computation, or the LPC residual signal analysis [44–46]. Due to the strong dynamics of the energy levels in the speech signal, and due to the difficult determination of the speech/nonspeech decision threshold, a new statistical-model-based voice activity detection strategy, slightly similar to the approach in [48], is applied in the proposed algorithm. In the first step, a preliminary additive noise reduction procedure is performed at the level $j = 5$ of the wavelet packet decomposition tree. Then, a denoised speech signal is reconstructed using wavelet packet reconstruction. In the second step, the VAD features are extracted and the two-stage statistical classifier is applied. In the first stage of the statistical classification, each frame m of the input signal is declared as speech or nonspeech. In the second stage, each speech frame is further declared as voiced or unvoiced. For voiced/unvoiced detection, a slightly modified feature vector is applied, as in the case of speech/nonspeech detection. The two statistical classifiers used in speech/nonspeech and voiced/unvoiced detections are based on Gaussian mixture models (GMM) [49]. The speech/nonspeech decision is used in the proposed primary noise reduction procedure. The voice/unvoiced decision is used in the adaptation process of the wavelet packet decomposition tree to extract the wavelet packet speech parameters. Under the presumption that energy-independent features are selected in the VAD procedure, the proposed VAD algorithm is robust against high variation of the input speech signal's energy. Furthermore, as GMM models are trained using speech data from many speakers, the proposed GMM-based voice activity detection procedure is robust against the speaker variability (speaking style, gender, age, etc.).

5.1. Feature vector definitions for speech activity and voicing detection

To achieve successful detection of speech frames in the input noisy speech signal using statistical classifier, discriminative features must be chosen, which enable good speech/nonspeech discrimination. The human speech production process can be mathematically well described by the usage of lower-dimensional autoregressive modeling [1, 2, 16]. Therefore, in the proposed statistical speech/nonspeech classification process, a feature vector composed of 10 linear predictive cepstral coefficients (LPCC) will be applied. These 10 LPCC coefficients will be computed using an autoregressive model of the order 12 [12, 50, 51]. In the voiced/unvoiced classification procedure, another voicing feature will be added to the proposed feature vector of 10 LPCC elements, composed only of a feature vector of 11 elements.

The preprocessed noisy input speech signal is denoised at the preliminary noise reduction stage using 5-level wavelet packet decomposition, the smoothed Donoho's threshold determination procedure, and the smoothed modified soft thresholding procedure. Then, the denoised signal is reconstructed. The 12-order autoregressive modeling is applied and 10 LPCC features are extracted for each frame m of the input speech signal. The vector of 10 LPCC elements is used in the speech/nonspeech classification procedure. The following paragraph describes the definition of the proposed voicing parameter ϑ, used as the 11th feature element in the feature vector for the voiced/unvoiced classification process.

An analytical sinusoidal model of speech signal production was presented in [46]. The analytical model of speech signal can be simplified into the following notation:

$$s[n] = \sum_{q=1}^{Q} A_q \cos\left[(n - n_0) q f_0 + \varphi_q\right], \tag{7}$$

where n_0 represents the speech onset time, Q is the number of harmonically related sinusoids with amplitudes of A_q and with phases φ_q. The fundamental frequency of the speech is denoted by f_0. The LPC residual error signal, denoted by $e[n]$, can be defined, using the following P-order inverse autoregressive (AR) filter:

$$e[n] = s[n] + \sum_{i=1}^{P} a_i s[n-i], \tag{8}$$

where $n = 0, 1, \ldots, N-1$ and $s[n] = 0$ if $n < 0$. The number of samples in the current frame is represented by N, and n presents the sample index in the frame m. On the basis of a simplified sinusoidal model of the speech signal, the following properties can be observed [46]: (1) the LPC residual signal of the stationary voiced speech is a deterministic signal, composed of Q sinusoids with equal amplitudes A_q, and harmonically related frequencies, (2) the LPC residual signal of the unvoiced speech can be represented as a harmonic process composed of Q sinusoids with randomly distributed phases φ_q.

The LPC residual signal of the noise with Gaussian distribution has the properties of the white Gaussian noise [46]. This important property of the LPC residual signal is used together with the well-known properties of higher-order cumulants. Namely, the cumulants of order c greater than 2 ($c > 2$) are equal to zero for the white Gaussian process [46]. In other words, higher-order cumulants are immune to white Gaussian noise. The primarily used higher-order cumulants are of the third order γ_3 (skewness) and fourth order (kurtosis) γ_4 cumulants, which are determined using the following notation:

$$\begin{aligned}
\gamma_3 &= E\{e^3[n]\} = \frac{1}{N}\sum_{n=0}^{N-1} (e[n])^3, \\
\gamma_4 &= E\{e^4[n]\} - 3\left[E\{e^2[n]\}\right]^2 \\
&= \frac{1}{N}\sum_{n=0}^{N-1} (e[n])^4 - 3\left[\frac{1}{N}\sum_{n=0}^{N-1} (e[n])^2\right]^2.
\end{aligned} \tag{9}$$

It was shown in [46], that the skewness γ_3, and the kurtosis γ_4 of the LPC residual signal depend only on the number of harmonically related components, and on the energy of the analyzed signal $s[n]$. The signal's energy influence on the voiced/unvoiced classification should be discarded. Therefore, the voicing parameter ϑ will be defined as an energy-eliminating ratio between the third (skewness) and fourth (kurtosis) order cumulants, which depend only on the number of harmonics Q in the analyzed speech signal [46]:

$$\vartheta = \frac{\gamma_3^2}{\gamma_4^{3/2}} = \frac{9(Q-1)^2}{8Q\left((4/3)Q - 4 + 7/6Q\right)^{3/2}}. \tag{10}$$

The above equation has a drawback, namely that it can become undetermined if the number of harmonics Q in the input signal is zero ($Q = 0$): this is the case when there is only a white Gaussian noise or unvoiced speech signal on the input. This condition rarely occurs due to variations in the cumulant estimates. Nevertheless, in the computation procedure, the following limitation is taken into account: if $Q = 0$, then the voicing parameter $\vartheta = 0$. The number of harmonics Q is computed by counting the local maxima of the LPC-based spectrum.

5.2. *Statistical classifier for speech activity and voicing detection*

A two-stage statistical classifier is applied in the proposed noise robust speech parameterization algorithm to perform speech/nonspeech and voiced/unvoiced classifications. Figure 6 shows a block diagram of the proposed two-stage statistical classifier. In the first stage, speech/nonspeech detection is performed for each frame m of the input signal. Then, in the second stage, each previously detected speech frame is further classified as voiced or unvoiced. The two statistical classifiers are based on the Gaussian mixture modeling (GMM) of input data. During the training phase, separate estimations of the speech and nonspeech frames were performed using the training part of the speech database. Similarly, the voiced and unvoiced GMM models were estimated. These four GMM models were then used to classify data from each new input signal frame. It was discovered that the usage of 32 continuous density Gaussian mixtures resulted in the best classification results. The training of GMM models was performed using the tools HInit (initial GMM parameter estimation using Viterbi algorithm), and HRest (implementation of the Baum-Welch iterative training procedure to find the optimal parameters of the GMM model with respect to the given input training data set), which are part of the HTK toolkit [49]. In the test phase, for each frame of the input signal, the emission probabilities of the corresponding GMM models are computed using the input feature vector. For example, if the voice activity detection of the frame m is performed, the speech and nonspeech GMM models are evaluated using the input LPCC feature vector of the frame m. As a result, two output log probabilities (called also emission probabilities in HMM-based ASR systems) are computed: $\log(\text{Prob}_{\text{SPEECH}}[m])$ and $\log(\text{Prob}_{\text{NONSPEECH}}[m])$. In the second stage, the voiced and unvoiced GMM models are evaluated for each speech-only frame of the input signal using corresponding feature vector (10 LPCCs + 1 voicing parameter ϑ). As a result of the second stage, the

```
    • First stage: voice activity detection $G[m]$:
$\forall m$, where $m$ is the input signal frame:
IF: $\log(\text{Prob}_{\text{SPEECH}}[m]) > \log(\text{Prob}_{\text{NONSPEECH}}[m])$
    THEN: $G[m] = 1$, the frame $m$ contains speech
ELSE: $G[m] = 0$, the frame $m$ does not contain speech

    • Second stage: voiced/unvoiced detection $Z[m]$:

Under condition $G[m] = 1$:
IF: $\log(\text{Prob}_{\text{VOICED}}[m]) > \log(\text{Prob}_{\text{UNVOICED}}[m])$
    THEN: $Z[m] = 1$, the frame $m$ contains voiced speech
ELSE: $Z[m] = 0$, the frame $m$ contains unvoiced speech
```

Algorithm 1

Table 1: The parameters of the WPD_1.

Level j	Output node index k
4	$8, 9, \dots, 15$
5	$8, 9, \dots, 15$
6	$0, 1, \dots, 15$
	The number of all output nodes: 32

Table 2: The parameters of the WPD_2.

Level j	Output node index k
4	$0, 1, \dots, 5$, and nodes $14, 15$
5	$12, 13, \dots, 17$, and nodes $26, 27$
6	$36, 37, \dots, 51$
	The number of all output nodes: 32

two log probabilities are computed: $\log(\text{Prob}_{\text{VOICED}}[m])$ and $\log(\text{Prob}_{\text{UNVOICED}}[m])$. The final binary classification results, $G[m]$ and $Z[m]$ are determined in Algorithm 1.

As evident, there is no need to define some special distance measure for speech/nonspeech and voicing classification: the two output probabilities of the GMM models are just simply compared to each other. Short pauses can often appear inside the spoken words in some cases. These short pauses usually appear before or after the stop phonemes, and can be misclassified as nonspeech segments. These misclassifications can decrease the performance of the automatic speech recognition system. To reduce the influence of possible fluctuations in the VAD output decision, the GMM emission log-probabilities $\log(\text{Prob}_{\text{X}}[m])$ are smoothed prior to generation of final decisions $G[m]$ and $Z[m]$. Smoothing is performed using the following first-order autoregressive lowpass filter:

$$\begin{aligned}\log\left(\text{Prob}_{\text{X}}[m]\right)' = {} & (1-\delta)\log\left(\text{Prob}_{\text{X}}[m]\right) \\ & + \delta\log\left(\text{Prob}_{\text{X}}[m-1]\right)'.\end{aligned} \tag{11}$$

The input speech data must be time labelled in order to train the GMM models. In the proposed procedure only the orthographic transcriptions were initially available. A forced Viterbi alignment procedure was applied to construct the corresponding time labels.

6. THE ADAPTIVE TOPOLOGY OF THE WAVELET PACKET DECOMPOSITION TREE

Many different possibilities exist for representing a speech signal in the time-frequency plane, by the usage of the wavelet packet decomposition. It is possible to select different wavelet-packet decomposition topologies, or various parameter sets [9, 10, 15, 20]. The proposed noise robust speech parameterization algorithm, WPDAM, exploits the advantages of the multiresolutional analysis provided by the wavelet packet decomposition of the speech signal. Furthermore, with the aim of improving the accuracy of the proposed speech representation in the time-frequency plane against the short time Fourier transform, the time and the frequency resolutions of the proposed speech signal analysis could be adapted to the characteristics of the speech signal. The basic speech units—phonemes—can be roughly divided into two main sets: voiced and unvoiced [1, 43]. It is already well-known that voiced speech is mainly concentrated in the low-frequency region, whereas the unvoiced speech has most of its spectral energy located at higher frequencies of the speech spectrum [43]. In the proposed WPD scheme the overall division of phonemes into the two main groups is exploited, as well as the spectral characteristics of both of them. The proposed WPD tree topology adaptation algorithm utilizes the output decision of the statistical voiced/unvoiced classifier $Z[m]$. On the basis of the two possible characterizations of the current speech frame m: frame m contains voiced speech if $Z[m] = 1$, or the frame m contains the unvoiced speech if $Z[m] = 0$, one of the two empirically determined wavelet packet decomposition tree topologies is selected:

$$\begin{aligned}&\text{IF}\quad Z[m] = 1\text{: the topology } \text{WPD}_1 \text{ is applied,}\\ &\text{IF}\quad Z[m] = 0\text{: the topology } \text{WPD}_2 \text{ is applied.}\end{aligned} \tag{12}$$

Figure 7 presents the definition of the WPD tree topology used to analyze voiced segments of the input speech signal. The wavelet packet parameters are calculated for the 32 output nodes of the corresponding 6-level wavelet packet decomposition tree. The relations between indexes k of the output nodes and corresponding decomposition levels j are represented in Table 1. The frequency resolution of the wavelet packet decomposition tree can be determined for each WPD level j using the following equation:

$$\Delta_f[j] = \frac{f_S}{2^{(j+1)}}, \tag{13}$$

where f_S represents the sampling frequency. Using the proposed WPD_1 topology, better frequency resolution at lower frequencies of the analyzed speech signal is achieved. Therefore, better description of the voiced segments of the speech signal is expected.

The opposite is true with the application of wavelet packet decomposition topology WPD_2, which is used to analyze unvoiced segments of the speech signal. The frequency

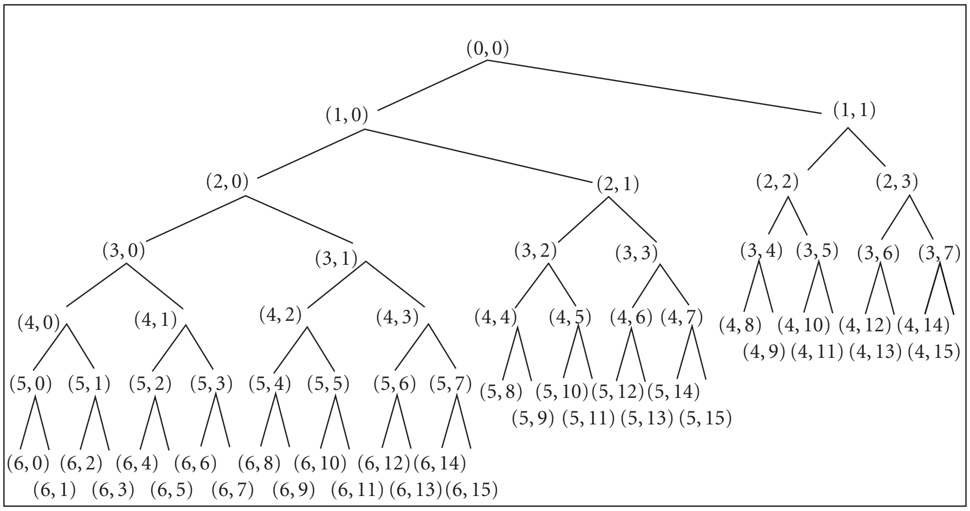

Figure 7: Topology WPD_1: voiced segments.

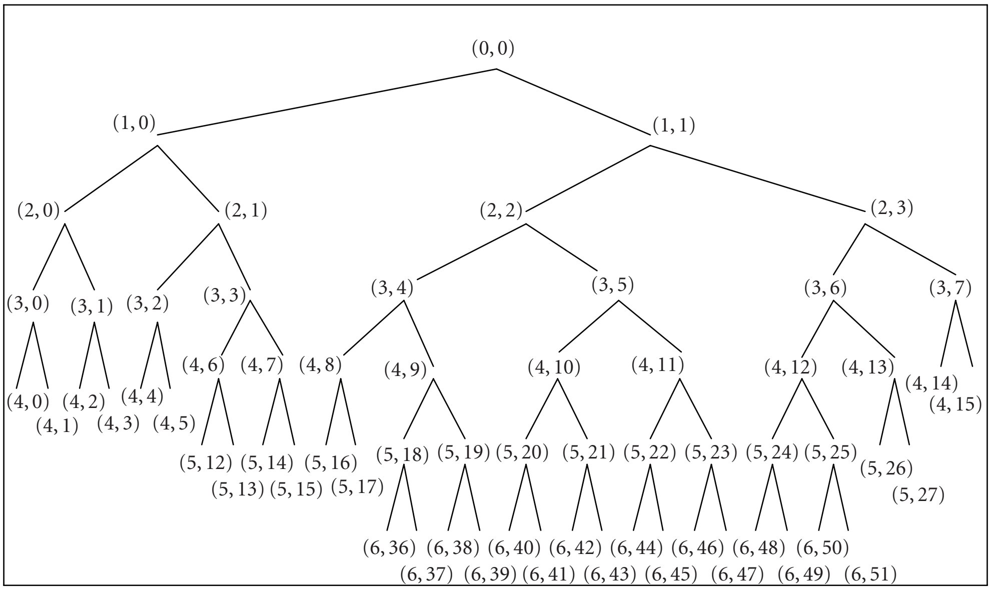

Figure 8: Topology WPD_2: unvoiced segments.

resolution at higher frequencies is increased and, therefore, the parameterization of the unvoiced segments of the speech signal is improved. The empirically defined wavelet packet decomposition tree topology WPD_2, used to analyze unvoiced segments of the speech signal, is represented in Figure 8. In this case the wavelet packet parameters are also computed for the 32 output nodes of the decomposition tree. The WPD_2 parameters are described in Table 2.

The presented optimal topologies WPD1 and WPD2 were determined with the analysis of average spectral energy properties of voiced and unvoiced speech segments of the studio quality database (TIDIGITS). This analysis shows for example that for unvoiced speech segments there is no benefit if nodes $(4, 14)$, $(4, 15)$, $(5, 26)$, and $(5, 27)$ are decomposed further (see Figure 8). Namely, it was discovered that the most important spectral region of majority of consonants is up to around 3400 Hz [2]. This frequency is also a bandwidth limit in the PSTN telephone network.

It should be noted that if the frame m does not contain any useful speech information (the VAD detection $G[m] = 0$), then it is discarded from further processing. This principle corresponds to the well-known frame dropping method [28].

7. WPD-BASED SPEECH PARAMETERS

A variety of different possibilities for selecting appropriate speech-describing features exist within the frame of WPD-based signal analysis [9, 10, 20]. In the proposed WPDAM, the basic WPD-based features will correspond to the energies of the wavelet packet sequences, computed on the terminal (output) nodes of the proposed wavelet packet decomposition tree (WPD1 or WPD2). The idea behind the usage of the energy as a main feature is motivated by the findings in the domain of psychoacoustics [1, 2, 16]. Namely, the fundamental speech processing information of human auditory system consists of the amount of speech energy located in the particular frequency subband [2]. The energies are computed with the application of the following equation:

$$E_k^j[m] = \frac{1}{N_k^j}\sum_{n=0}^{N_k^j-1}\left(W_k^j(s'[m,n])\right)^2, \tag{14}$$

where $W_k^j(s'[m,n])$ denotes wavelet packet decomposition sequence of the node k on the level j, and of the length of N_k^j, computed for the noise-reduced speech signal $s'[m,n]$. Therefore, the computed energy parameters $E_k^j[m]$ represent the results of the WPD-based multiresolutional speech signal analysis.

7.1. *The combined root-log compression characteristics*

The logarithmic (log) compression characteristic is the most frequently used parameter compression mode in speech parameterization algorithms to reduce the dynamics of parameters (like filterbank energies compression prior to the DCT calculation in the MFCC extraction procedure) [7]. Nevertheless, some authors reported about the usage of exponential (or root) compression characteristics instead of log compression, and achieved better automatic speech recognition performance under noisy conditions [52]. In the presented noise robust speech parameterization algorithm the combined root-log compression characteristics is proposed as:

$$P_k^j[m] = \begin{cases} \sqrt[r]{E_k^j[m]}, & \text{if } E_k^j[m] < B, \\ \log\left(E_k^j[m]\right), & \text{if } E_k^j[m] \geq B. \end{cases} \tag{15}$$

If the value of the energy $E_k^j[m]$ is lower than the value of the predefined breakpoint B, then the root compression of the degree r is used, otherwise the logarithmic compression characteristic is applied. The values of r and B must be appropriately determined in (15) in order to achieve smooth contour of the compression characteristic. First, the value of r is selected with respect to the condition that the two (root and log) characteristics are intersecting exactly at the breakpoint B:

$$\begin{aligned} \log B = x \wedge \sqrt[r]{B} = x &\longrightarrow \log B = B^{1/r}, \\ &\longrightarrow r = \frac{\log(B)}{\log\left(\log(B)\right)}. \end{aligned} \tag{16}$$

The breakpoint B is set to 1% of the maximum value of the uncompressed wavelet packet energy parameter $E_k^j[m]$, determined on the basis of the training part of the speech database (Aurora 2).

8. PRIMARY FEATURE VECTOR BASED ON JOINT WPD AND AUTOREGRESSIVE MODELING

The primary feature vector $\mathbf{x}[m]$, proposed in the presented noise robust speech parameterization procedure, is defined as a concatenation of 10 linear predictive cepstral coefficients (LPCC) $\mathbf{a}_{\text{LPCC}}[m]$, and of the 33 root-log compressed parameters of the wavelet packet decomposition $P[m]$:

$$\mathbf{x}[m] = \{\mathbf{a}_{\text{LPCC}}[m], P[m]\}. \tag{17}$$

The autoregressive modeling of speech signal provides a good description of the speech production system and encompasses information about the spectral envelope of the speech signal. These low-dimensional parameters are especially well-defined if the speech signal is voiced and periodic (quasiperiodic) [11, 12, 50, 51]. The most important discriminant information about voiced phonemes is the shape of the spectral envelope, as well as the position and the magnitude of the particular formant [1, 12, 50].

The speech information contained in the LPCC coefficients $\mathbf{a}_{\text{LPCC}}[m]$ can be well supplemented with the information carried by the wavelet packet parameters $P[m]$. Namely, the WPD provides multiresolutional analysis of the speech signal and, therefore, the parameters $P[m]$ also provide a good description of unvoiced segments of the speech signal (see Figure 8). The parameters, constructed on the basis of the autoregressive model are combined together with wavelet packet decomposition parameters, to build a primary feature vector, which provides better parameterization of speech signal than the separate use of the above-mentioned two parameterization modes.

The primary feature vector $\mathbf{x}[m]$, constructed using the proposed noise robust speech parameterization algorithm, contains 43 elements in total: there are 10 LPCC parameters $\mathbf{a}_{\text{LPCC}}[m]$, already computed in the voice activity detection stage (see Section 5.1), as well as 33 root-log compressed wavelet packet decomposition parameters $P[m]$. Before the feature vector postprocessing procedure is carried out, the primary feature vector $\mathbf{x}[m]$ is supplemented with dynamical coefficients—with its first $\mathbf{\Delta}[m]$ and second order $\mathbf{\Delta\Delta}[m]$ derivatives [49].

9. FEATURE VECTOR POSTPROCESSING PROCEDURE

The output of the previous processing stages of the WPDAM is a primary feature vector of the total length of 129 elements (3 × 43 elements). The feature vector postprocessing procedure is applied to simultaneously reduce the dimension of the final output feature vector, and to enhance the performance of the classification process. The main tasks and goals of the proposed feature vector postprocessing procedure are the following [1, 18, 53].

(i) To reduce the acoustical mismatch between the training and testing environments.

(ii) Reduction in the dimensionality of the feature vectors and, therefore, the increase in accuracy of the acoustical modeling [18]. Nevertheless, with the usage of lower-dimensional features, the computational and memory requirements are also reduced.

(iii) To increase the discriminativity between different classification data classes (phonemes). This leads to immediate enhancement of the automatic speech recognition accuracy.

(iv) Decorrelation of the feature vectors' elements, which enables the usage of the diagonal covariance matrices in the hidden Markov modeling process. The usage of diagonal covariances also reduces the computational and memory requirements of the automatic speech recognition system [14, 49].

9.1. Acoustic mismatch reduction between the training and testing environments using a statistics-based transformation

Automatic speech recognition algorithms based on hidden Markov models assume that the training and testing materials correspond to the same data distribution. Increase in the acoustical mismatch between the training and testing environments proportionally reduces the classification accuracy [4]. The main purpose of the presented acoustic mismatch reduction procedure is the definition of the transformation, which will assure similarity between the training and testing data distributions. To simplify the problem, the simple Gaussian data distribution will be assumed in the following derivations. Furthermore, contamination by additive noise is assumed:

$$\mathbf{x} = \mathbf{s} + \boldsymbol{\eta}. \tag{18}$$

In (18) $\mathbf{s}$ denotes the feature vector of the clean speech, $\boldsymbol{\eta}$ is the feature vector of additive noise, and $\mathbf{x}$ represents the feature vector of the noisy speech signal. Assuming that the vectors $\mathbf{s}$ and $\boldsymbol{\eta}$ correspond to the Gaussian data distribution, then the estimated clean-speech feature vector $\mathbf{a}$ can be derived from the feature vector of the noisy speech $\mathbf{x}$ using the following equation:

$$\mathbf{a} = \beta(\mathbf{x}) + \alpha = \beta(\mathbf{s} + \boldsymbol{\eta}) + \alpha, \tag{19}$$

where α (biasing factor) and β (scaling factor) are free variables that need to be determined. The feature vectors $\mathbf{a}$ and $\mathbf{s}$ should correspond to the same Gaussian data distribution: their mean values and variances must be equal:

$$\overline{\mathbf{a}} = \overline{\mathbf{s}}, \qquad \sigma_{\mathbf{a}}^2 = \sigma_{\mathbf{s}}^2. \tag{20}$$

The simple Gaussian distribution is completely determined by its mean value and variance. It will be shown that (19) is valid, if the free variables α and β are determined as

$$\beta = \frac{1}{\sqrt{1 + (\sigma_{\boldsymbol{\eta}}^2/\sigma_{\mathbf{s}}^2)}}, \qquad \alpha = \overline{\mathbf{s}}(1 - \beta) - \beta\overline{\boldsymbol{\eta}}. \tag{21}$$

Firstly, the variance of the feature vector $\mathbf{a}$ will be defined using the following equation:

$$\begin{aligned}(\mathbf{a} - \overline{\mathbf{a}})^2 &= \left(\beta(\mathbf{s} + \boldsymbol{\eta}) - \beta(\overline{\mathbf{s}} + \overline{\boldsymbol{\eta}})\right)^2 \\ &= \beta^2\left((\mathbf{s} - \overline{\mathbf{s}}) + (\boldsymbol{\eta} - \overline{\boldsymbol{\eta}})\right)^2 \\ &= \beta^2\left((\mathbf{s} - \overline{\mathbf{s}})^2 + 2(\mathbf{s} - \overline{\mathbf{s}})(\boldsymbol{\eta} - \overline{\boldsymbol{\eta}}) + (\boldsymbol{\eta} - \overline{\boldsymbol{\eta}})^2\right).\end{aligned} \tag{22}$$

The speech signal $\mathbf{s}$ and the additive noise $\boldsymbol{\eta}$ are uncorrelated, therefore, (22) can be written in the following form:

$$\sigma_{\mathbf{a}}^2 = \beta^2(\sigma_{\mathbf{s}}^2 + \sigma_{\boldsymbol{\eta}}^2). \tag{23}$$

If the properties of (20) are taken into consideration, the variable β can be defined as

$$\begin{aligned}\sigma_{\mathbf{a}}^2 = \sigma_{\mathbf{s}}^2 = \beta^2(\sigma_{\mathbf{s}}^2 + \sigma_{\boldsymbol{\eta}}^2) \Longrightarrow \beta &= \frac{\sigma_s}{\sqrt{\sigma_{\mathbf{s}}^2 + \sigma_{\boldsymbol{\eta}}^2}} \\ &= \frac{1}{\sqrt{1 + \sigma_{\boldsymbol{\eta}}^2/\sigma_{\mathbf{s}}^2}}.\end{aligned} \tag{24}$$

The variable α is defined using (19) and (20):

$$\overline{\mathbf{s}} = \overline{\mathbf{a}} = \beta(\overline{\mathbf{a}} + \overline{\boldsymbol{\eta}}) + \alpha \Longrightarrow \alpha = \overline{\mathbf{s}}(1 - \beta) - \beta\overline{\boldsymbol{\eta}}. \tag{25}$$

The mean values and variances $\overline{\mathbf{s}}$, $\overline{\boldsymbol{\eta}}$, $\sigma_{\mathbf{s}}^2$, and $\sigma_{\boldsymbol{\eta}}^2$, used to determine the parameters α and β were estimated using the Aurora 2 database, where both the speech and noisy versions of the same utterances are available. The estimation of α and β is performed only once using the Aurora 2 noisy speech utterances with the SNRs of 10 dB. Once estimated, the two α and β parameters were used in all other cases (including the Aurora 3 database).

9.2. Linear discriminant analysis (LDA)

Additionally, linear discriminant analysis (LDA) is applied to transform the higher dimensional feature vector $\mathbf{a}[m]$ to the lower-dimensional final output feature vector $\mathbf{b}[m]$, and to simultaneously increase the centroid distances of the K discrimination classes. This is in order to reduce the computational load of the automatic speech recognition system and to enhance the classification process. The basic idea of LDA is to reduce the variances within the classes, whereas the variances between the classes should be as large as possible [54]. The following steps describe the LDA processing procedure.

Determination of LDA classes

In the proposed LDA procedure [53], the K classes correspond to the emitting states of the hidden Markov models (HMM) for all phonemes in the dictionary [29]. The only exceptions are /sp/ and /sil/: the short pause and silence models, respectively. In the first step, initial 16 Gaussian-mixture monophone HMMs, each with 3 emitting states, are trained using the clean-speech training data of the Aurora 2 database and original input feature vectors $\mathbf{a}[m]$ of length 129 elements. Then, the forced Viterbi full state alignment procedure [49, 53] was applied to divide input feature vectors $\mathbf{a}[m]$ of all the training materials into the K classes.

LDA transformation

Once the K classes are determined, the LDA transformation matrix $\mathbf{\Omega}$ is computed using the procedure defined in [53]. After taking into account the subtraction of the global mean value (mean feature vector $\mathbf{m}$), the final output feature vector is computed using LDA transform:

$$\mathbf{b}[m] = \mathbf{\Omega}^T(\mathbf{a}[m] - \mathbf{m}). \tag{26}$$

The dimensionality of the final output feature vector is reduced using the above-described procedure from the initial 129 elements (feature vector $\mathbf{a}[m]$) to 39 elements ($\mathbf{b}[m]$). This feature vector dimension is most commonly used in many of modern ASR systems. The final output feature vectors $\mathbf{b}[m]$, produced using the presented WPDAM, are used in the acoustic HMM training procedure, as well as in the robust ASR performance evaluation procedures.

10. EXPERIMENTAL FRAMEWORK AND RESULTS

Connected digit recognition experiments were performed using the Aurora 2 [29] and Aurora 3 [31–34] databases, which were designed to evaluate the performance of feature extraction algorithms under different noisy and acoustic mismatch conditions. The standard training and testing procedures that have been specified by the Aurora group for standardizing distributed speech recognition (DSR) front end were used together with the HTK hidden Markov model toolkit [30, 35, 49]. The whole word acoustic models are composed of 16 emitting states, each of 3 mixtures per state (with the exception of the silence model which used 3 emitting states and 6 mixtures per state) [30, 35]. The automatic speech recognition performance of the WPDAM was evaluated by a comparison with the standard baseline MFCC front ends, which were also determined by the Aurora DSR group [29, 30, 35].

Description of Aurora 2 and Aurora 3 databases and experiments

Aurora 2: The speech data in Aurora 2 database [29] is a derivative of the TI-DIGITS database. 8440 utterances (connected digits) were chosen for clean-condition training. On the other hand, for the multiconditional training, the same 8440 utterances were divided into 20 subsets with 422 utterances. The 20 subsets represented 5 different noise scenarios (suburban train, babble, car, exhibition noise, and clean scenario). These noises were added to each subset at SNRs of 20, 15, 10, and 5 dB. Three different test sets were defined to simulate the matched acoustic condition (set A), mismatched acoustic condition (set B), and the mismatched channel condition (set C). 4004 digit strings were first selected from the test part of TI-DIGITS, and then four subsets with 1001 utterances were obtained.

The test set A consists of 28028 utterances obtained by the addition of four types of noises (the same noises as in the multiconditional training procedure) at SNRs 20, 15, 10, 5, 0, −5 dB, and clean (no noise) condition to all subsets. There was a high match between test set A and the multiconditional training procedure because this test set used the same noises as multiconditional training.

Table 3: Baseline ETSI ES 201 108 absolute overall Aurora 3 performance (% accuracy).

Condition	ETSI ES 201 108 MFCC Baseline (% acc.)
WM	91.04%
MM	78.05%
HM	51.16%
Overall ACC	76.52%

The second test set, set B, was constructed in a similar way to set A. The only difference was that there were four different kinds of noise (restaurant, street, airport, and train station). Therefore, there was a mismatch between the training and testing conditions.

The third test set, set C, was created to simulate a mismatched channel characteristic. Set C contained 2 out of 4 subsets each with 1001 utterances. Speech and noise (suburban train, and street) were filtered by a MIRS filter before being added to result in SNRs of ∞, 20, 15, 10, 5, 0, and −5 dB. MIRS is a frequency response that simulates the characteristic of narrowband telephone terminal devices [33].

Aurora 3: The four languages (subsets), Finnish, Spanish, German, and Danish of Aurora 3 databases are taken from the corpora recorded as part of the SpeechDat-Car project [31–34]. These are real-condition recordings recorded whilst driving cars, using a close-talking microphone and a hands-free microphone. Three train/test configurations were defined for each of the four subsets separately: the well-matched condition (WM), the medium mismatched condition (MM), and the highly mismatched condition (HM). In the WM case, 70% of the entire data was used for training and the remaining 30% for testing. Therefore, the training test contained all the variability that appears in the test set. In the MM case, only hands-free microphone data was used for both training and testing. In the HM case, the training data consisted of close microphone recordings while testing was performed using distance-talking microphone data. Table 3 presents the Aurora 3 baseline absolute ASR performance of the standardized MFCC feature extraction procedure ETSI ES 201 108 [5].

10.1. Separate evaluation of particular WPDAM processing steps

The proposed noise robust speech parameterization algorithm WPDAM consists of several processing steps, which all contribute to the overall automatic speech recognition performance of the WPDAM. In this section, the performance contributions of certain particular processing steps are evaluated separately using the Aurora 3 database. The remaining processing steps, which are not in focus during the particular experiment, were kept unchanged for evaluation purposes.

Speech signal preprocessing

Comparison between the automatic speech recognition results shows that the proposed speech parameterization procedure, which consists of high-pass and pre-emphasis filtering, produces an 0.82% higher average absolute performance than in the case where no preprocessing is applied. It is also noticeable that classical first-order pre-emphasis [5] produces 0.46% lower absolute performance than the proposed preprocessing procedure. Additionally, it was shown that the proposed pre-emphasis procedure has greater influence on the automatic speech recognition performance than high-pass filtering.

Wavelet packet decomposition

The presented noise robust speech parameterization algorithm WPDAM uses a wavelet packet decomposition to represent the speech signal in the time-frequency plane. The performances of the two different filter types were compared, namely, the well-known Daubechies wavelet DB16, and the proposed finite impulse response filter REMEZ32 constructed on the basis of the Parks-McClellan procedure (Remez exchange algorithm [16, 41]). The results prove the hypothesis that the most important properties of the decomposition filters applied in the WPD-based speech parameterization procedure are good separability between pass- and stop-bands (narrow transition band), and relative high attenuation in the stop band. Namely, the proposed REMEZ32 filter has the steepest transit from the pass- to stop band when compared to the DB16. The DB16 has a wider transition band, and, therefore, worse frequency separability, resulting in the DB16 producing the lowest automatic speech recognition performance. Nevertheless, the DB16 also enables perfect reconstruction and is therefore very useful for coding purposes, but when used in feature extraction, it causes frequency component mixing due to its gradual transit from the pass to stop band.

WPD based additive noise reduction

As described in Section 4, the level of additive noise can be reduced effectively in the wavelet packet decomposition domain using thresholding techniques with time-frequency adaptive threshold. It can be stated that the continuity of the thresholding function has an important impact on the quality of the WPD-based denoised speech signal. Namely, it can be observed that in the case of the classical hard thresholding technique a 2.37% lower average absolute speech recognition performance is achieved than in the case of classical soft thresholding technique. The impact of the threshold determination technique was also investigated. When using the time-frequency adaptive threshold, a 2.49% higher average absolute performance is achieved than with the usage of a universal Donoho's threshold and modified smoothed soft thresholding (compare WPDAM and UT-SMT). Namely, the speech signal is a very dynamic process with highly nonstationary frequency content and, therefore, the threshold used in the noise reduction algorithm must be adaptive at the time—as well as in the frequency dimension.

Voice activity and voiced/unvoiced detection

The accurate voice activity and voiced/unvoiced detections are essential processing steps in the WPDAM. Namely, the two output decisions are applied to estimate the parameters of the denoising procedure, as well as to perform the WPD topology adaptation. The automatic speech recognition results show that the proposed GMM-based statistical VAD classifier achieves a 0.56% better average absolute performance than the energy-based VAD, defined in the standard ETSI ES 202 050 [6]. The differences in the performances of both VAD approaches become increasingly noticeable under medium and highly acoustical mismatched conditions. The usage of the proposed GMM-based voiced/unvoiced classification procedure achieves a 1.48% higher absolute average performance than in the case of pure energy based voiced/unvoiced decision. It can be stated that energy-based approaches are very sensitive to increasingly acoustic mismatching conditions.

Adaptive topology of the wavelet packet decomposition tree

The highest automatic speech recognition performance is achieved in the case of adaptive WPD topology. Using only the fixed topology WPD_1 a 6.11% better absolute automatic speech recognition performance is achieved than with application of the fixed topology WPD_2 only. The reason for this lies in the nature of the speech itself, which contains higher amounts of voiced phonemes than unvoiced. Namely, the topology WPD_2 accommodates the analysis of unvoiced speech. Therefore, voiced speech is, in the process of acoustic modeling, statistically more important than the unvoiced. But this fact should not be misinterpreted. The accurate identification and recognition of unvoiced speech is also very important, in order to achieve higher automatic speech recognition performances. Namely, this is the main advantage of the proposed noise robust speech parameterization algorithm WPDAM, which uses adaptive topology of the wavelet packet decomposition tree, and voiced/unvoiced detection. Therefore, the voiced, as well as unvoiced speech-segments can be accurately analyzed.

The proposed combined root-log compression characteristics

The results show that the proposed combined root-log compression achieves the best automatic speech recognition performance when compared to independently used root or log compression characteristics. Namely, the combined root-log compressed parameters reflect higher signal-to-noise ratio (SNR) as separate root or log compressed WPD parameters [52].

Table 4: Aurora 3 performance evaluation of WPDAM.

Subset Condition	Finnish acc. %	Spanish acc. %	German acc. %	Danish acc. %	Avg. acc.	Rel. improv.
WM	95.77	95.61	94.93	92.71	94.76	41.16
MM	90.03	93.02	89.24	78.95	87.81	46.47
HM	86.85	90.12	90.15	73.53	85.16	69.28
Overall accuracy	91.53	93.33	91.74	83.10	89.93	—
Relative improv.	53.27	55.39	47.92	43.62	—	50.05%

Table 5: Aurora 2 performance evaluation of WPDAM: multiconditional training.

Aurora 2 multiconditional training—Absolute performance (% ACC) and relative improvement (% REL) to the ETSI ES 201 108															
Set	Test set A					Test set B					Test set C				% REL
SNR	Subway	Babble	Car	Exhibition	Avg. acc.	Restaurant	Street	Airport	Station	Avg. acc.	Subway	Street	Avg. acc.	Overall acc.	Rel. improv.
Clean	99.20	98.97	98.93	99.20	99.08	99.20	98.97	98.93	99.20	99.08	99.08	98.97	99.03	99.07	30.11
20	98.80	98.67	98.81	98.58	98.72	98.43	98.22	98.42	98.86	98.48	98.50	98.22	98.36	98.55	36.03
15	97.88	97.91	98.51	97.38	97.92	98.00	97.61	97.73	97.75	97.77	97.76	97.46	97.61	97.80	32.29
10	95.79	95.84	96.69	95.23	95.89	94.87	95.56	95.41	96.02	95.47	95.58	95.25	95.42	95.62	21.84
5	91.25	89.00	91.80	90.02	90.52	86.71	90.42	89.47	90.19	89.20	91.00	89.00	90.00	89.89	24.73
0	78.63	68.23	78.26	74.09	74.80	65.28	74.94	74.89	73.74	72.21	74.55	73.88	74.22	73.65	32.77
−5	46.58	32.83	42.38	41.97	40.94	30.49	43.05	41.31	41.44	39.07	40.96	39.90	40.43	40.09	20.10
Avg.	92.47	89.93	92.81	91.06	91.57	88.66	91.35	91.18	91.31	90.63	91.48	90.76	91.12	91.10	—
Rel. improv.	32.13	16.54	33.57	25.41	26.91	25.69	25.63	27.05	39.35	29.43	39.88	30.04	34.96	—	29.53%

Feature vector definition based on joint wavelet packet decomposition and autoregressive modeling

The primary feature vector is in the proposed noise robust speech parameterization procedure WPDAM, constructed using a combination of autoregressive parameters and compressed wavelet packet decomposition parameters. With the proposed primary feature vector, better automatic speech recognition (89.93%) is achieved than in those cases where the autoregressive parameters (89.59%) and compressed wavelet packet decomposition parameters (89.69%) are used separately, and independently. Both complementary speech parameterizations, therefore, together enable a better description of the information contained in the speech signal and, thus, also higher automatic speech recognition accuracy.

Feature vector postprocessing

The last processing step applied in the WPDAM is feature vector postprocessing. It was shown that the proposed statistical modeling procedure decreases the average speech recognition accuracy (−0.25%) in the case of well-matched conditions, but in the case of medium-mismatched or highly mismatched acoustic conditions, the automatic speech recognition is increased (+0.49% and +1.21%, resp.). An automatic speech recognition performance reduction of 0.58% is observed with the application of PCA (principal component analysis) instead of LDA. Better automatic speech recognition performance is achieved with the application of LDA, due to its better class discriminability.

10.2. WPDAM Aurora 3 performance evaluation

Table 4 shows the absolute automatic speech recognition results ACC[%], achieved using the proposed WPDAM on the Aurora 3 speech database. The same table also presents the relative performance improvements against the baseline reference system based on ETSI ES 201 108 [5]. An average overall absolute improvement of 13.41% is achieved with the WPDAM procedure when compared to the standardized MFCC procedure. The achieved average relative performance improvement of the baseline system is 50.05%. Higher relative improvements are achieved under higher mismatched conditions. The average relative improvement under high-mismatched conditions is 69.28%, and under well-matched condition 41.16%. In the case of the German part of the Aurora 3 database, interesting results are evident: in the case of the highly mismatched condition, better automatic speech recognition performance than in the case of the medium-mismatched condition can be observed (ACC = 90.15% in HM condition versus ACC = 89.24% in MM condition). It is assumed that the observed anomaly originates in the reduced statistical reliability of the German part of Aurora 3, due

Table 6: Aurora 2 performance evaluation of WPDAM: clean training.

Aurora 2 clean training—Absolute performance (% ACC) and relative improvement (% REL) to the ETSI ES 201 108															
Set	Test set A					Test set B					Test set C				% REL.
SNR	Subway	Babble	Car	Exhibition	Avg. acc.	Restaurant	Street	Airport	Station	Avg. acc.	Subway	Street	Avg. WER	Overall WER	Rel. improv.
Clean	99.36	99.24	99.19	99.32	99.28	99.36	99.18	99.19	99.47	99.30	99.32	99.26	99.29	99.29	22.45
20	98.10	97.79	98.39	98.14	98.11	97.64	97.49	98.09	98.33	97.89	97.96	98.11	98.04	98.00	60.89
15	95.15	96.16	97.05	95.65	96.00	95.09	96.19	96.75	96.73	96.19	96.47	96.83	96.65	96.21	72.39
10	89.47	89.15	93.56	88.86	90.26	88.61	91.78	92.93	92.99	91.58	93.01	92.75	92.88	91.31	74.70
5	76.85	74.09	84.25	73.71	77.23	72.44	80.64	82.77	81.95	79.45	83.72	83.69	83.71	79.41	67.56
0	54.03	44.36	58.78	51.66	52.21	45.56	55.88	59.20	59.39	55.01	59.83	58.92	59.38	54.76	46.98
−5	31.69	24.32	33.38	32.35	30.44	20.40	27.42	28.01	27.99	25.96	28.57	29.23	28.90	28.34	22.94
Avg.	82.72	80.31	86.41	81.60	82.76	79.87	84.40	85.95	85.88	84.02	86.20	86.06	86.13	83.94	—
Rel. improv.	47.60	69.92	66.51	53.54	59.39	67.37	59.85	77.24	73.19	69.41	66.57	63.26	64.92	—	64.50%

to the relatively small training set defined for the medium-mismatched condition.

10.3. WPDAM Aurora 2 performance evaluation

Table 5 shows the absolute automatic speech recognition accuracy achieved using the proposed WPDAM procedure on the Aurora 2 speech database, with a multiconditional training procedure. It is evident from the table that, in the cases of speech-alike noises such as babble and restaurant, the lowest average speech recognition accuracy is achieved (ACC = 89.93% and 88.66%, resp.). It is commonly agreed that the competing speech noise represents, due to the highly nonstationary nature of the speech signal, one of the most difficult noise scenarios [39]. Namely, background speech has fast-changing spectral content, which is also very similar to the primary speech signal [1, 39]. It should be noted that the influence of competing speech is very hard to eliminate in one-channel systems. Even a human would have a major problem listening and understanding low-SNR babble speech using one ear only [2]. When using the WPDAM an overall absolute Aurora 2 multiconditional performance improvement of 4.07% is achieved, when compared to the standardized MFCC feature extraction procedure ETSI ES 201 108 [5]. It is also evident from Table 5 that, in the case of multiconditional training, a relative improvement of 29.53% is achieved in comparison to the baseline system. The partial results show the highest relative improvements for highly mismatched test sets C (REL = 34.96%) and B (REL = 29.43%). As expected, the lowest improvement is achieved under babble noise condition (REL = 16.54%). This result also confirms the assumption that the competing speaker noise represents one of the most difficult noisy conditions. Similar critical conditions are exhibition, street, and restaurant. In opposition to the above-mentioned noise types, the following are less critical and, therefore, better relative improvements can be achieved with them: station (REL = 39.35%), subway (REL = 32.13%), and car (REL = 33.57%).

Table 6 shows the automatic speech recognition results of the WPDAM using the Aurora 2 database with clean training procedure. The acoustic mismatch between the training and testing environments is, in this case, very high. Therefore, this experiment shows a higher performance improvement by the WPDAM against the baseline feature extraction procedure ETSI ES 201 108, as an experiment using multiconditional training. An absolute overall improvement of 25.88% is observed. Nevertheless, the overall performance of the proposed WPDAM is, in the case of clean training (83.94%), still lower than in the case of multiconditional training (91.10%). However, the difference between the two training modes is, when using the WPDAM, smaller (7.16%) than the difference between the two training modes achieved with the standardized algorithm ETSI ES 201 108 (28.97%). This result proves that when using the WPDAM, the acoustic mismatch between the training and testing environments is efficiently reduced. Table 6 also shows the relative improvement achieved using the WPDAM on the Aurora 2 database, and with a clean training procedure. Much higher relative performance improvements are observed due to higher initial mismatch between training and testing environments. The total overall relative improvement of 64.50% is achieved when using the proposed noise robust speech parameterization procedure. The best partial result is achieved for the test set B (REL = 69.41%). It can also be observed that the average relative improvement increases with respect to the degrading of the SNR. It reaches its maximum at SNR = 10 dB (REL = 74.70%). With further degradation of the SNR, the average relative improvement decreases gradually and achieves 22.94% at the SNR = −5 dB.

10.4. Performance comparison of WPDAM against ETSI ES 202 050 (AFE)

In order to enable a performance comparison between the proposed noise robust feature extraction algorithm WPDAM and any other existing front ends in literature, it is sufficient to provide a comparison of the proposed algorithm against standardized feature extraction algorithms. The performance

Table 7: Aurora 3 performance evaluation of AFE.

Subset	Finnish acc. %	Spanish acc. %	German acc. %	Danish acc. %	Avg. acc.	Rel. improv.
Condition						
WM	95.91	96.74	95.15	93.44	95.31	47.70
MM	88.78	93.86	88.49	82.01	88.29	47.47
HM	86.25	91.82	90.93	79.47	87.12	73.08
Overall accuracy	91.00	94.50	91.76	85.95	90.80	—
Relative improv.	51.53	64.43	48.26	51.64	—	53.97%

Table 8: Aurora 2 performance evaluation of AFE.

Relative improv. [%]	Test set A	Test set B	Test set C	Overall relative impr.
Multi	28.37	36.79	32.68	32.60
Clean	70.32	74.74	64.17	70.86
Average	49.34	55.76	48.43	51.73%

comparison relative to the first Aurora MFCC standard ETSI ES 201 108 [5] has already been presented in this section (Tables 4–6). Motorola, France Telecom, and Alcatel proposed and also standardized advanced front-end (AFE) algorithm ETSI ES 202 050 v1.1.3 [6]. The AFE contains an improved noise reduction stage using a two-stage mel-warped Wiener filtering technique, and energy-based voice activity detection procedure. The baseline automatic speech recognition performance of the AFE procedure using Aurora 2 and Aurora 3 databases is given in [6]. The direct comparison about performances of the WPDAM and AFE on the Aurora 3 database can be seen from Tables 4 and 7. Tables 5, 6, and 8 present the comparison between WPDAM and AFE on the Aurora 2 database. When compared to AFE, the proposed WPDAM achieves 4.71% lower overall relative improvement on the Aurora 2 database, and 3.92% lower overall relative improvement on the Aurora 3 database, with respect to the baseline standard ETSI ES 201 108 [5]. However, WPDAM achieves better ASR performance at higher SNRs (> 20 dB), as well as with the test set C condition (channel mismatch) of the Aurora 2 noisy speech database.

10.5. WPDAM computational complexity and real-time deployment feasibility

Table 9 presents the results of a computational cost comparison (real time factor RTx) for the standardized feature extraction algorithms ETSI ES 201 108, ETSI ES 202 050, as well as the computational complexity of WPDAM. The real time factor RTx presents the needed processing time (in seconds) to process 1 second of input speech signal. The tests were performed on a Pentium 4-based (3 GHz) PC with hyperthreading functionality enabled. WPDAM is found to be 5.9 times slower than ETSI ES 201 108, and 2.5 times slower than the advanced front end ETSI ES 202 050. However, 15 feature extraction processes WPDAM can still be operated simultaneously with real-time operation. It should be mentioned, that in WPDAM implementation, no special care to code optimization has been performed currently.

11. CONCLUSION

This article presents a novel noise robust speech parameterization procedure WPDAM based on wavelet packet decomposition. ASR performance evaluation using the Aurora 3 database shows the efficiency and contribution of the particular processing step to the overall performance of the presented WPDAM. Finally, the ASR robustness experiments performed based on Aurora 2 and Aurora 3 noisy speech databases show an overall relative performance improvement of 47.02% and 50.05%, respectively (see Tables 3–6), relative to the baseline MFCC front-end ETSI ES 201 108 [5]. The AFE achieves of 4.71% higher Aurora 2, and of 3.92% higher Aurora 3 relative overall improvement when compared to the presented WPDAM algorithm (see Tables 7, 8). Nevertheless, in comparison to the AFE (ETSI ES 202 050), the WPDAM enables higher Aurora 2 performance for SNRs higher than 20 dB, as well as of 1.51% higher relative performance for mismatched channel condition (Aurora 2, test set C). WPDAM has proved to be robust to different noise characteristics, SNRs, and under various levels of acoustical matching between the training and testing conditions. Despite the fact that WPDAM underperforms the AFE in some aspects, it is still important from another point of view. It can be stated from the presented ASR results that similar or, in some cases even better, noise robust automatic speech recognition performance can be achieved with WPDAM, when compared to the generally known and used STFT, or warped Wiener filter-based approaches (ETSI ES 201 108 [5], ETSI ES 202 050 [6]). The short-time Fourier transform, introduced by Gabor in 1946 [13], dominates in the domain of time-frequency representation of a speech signal. Nevertheless, the results of the proposed research work coincide with the findings of other researchers in the domain of wavelet transform, and

Table 9: WPDAM computational complexity evaluation.

Feature extraction method	Real-time factor (RTx)	Number of parallel systems (1/RTx)
ETSI ES 201 108	0.0108	92
ETSI ES 202 050	0.0251	39
WPDAM	0.0637	15

wavelet packet decomposition [8–10, 15–27]. It has also been established that wavelet-based multiresolutional approaches have many advantages against the STFT based approaches. The presented work also opens up several possibilities for using the proposed ideas and algorithms, independently, in different speech processing areas. The proposed denoising approach can be used, for example, in the speech enhancement stage of a speech telecommunication system in order to improve the SNR of those speech signals captured in adverse environments. Similar alternative usage can also be found for the presented GMM-based voice activity detection and voiced/unvoiced detection procedure. Nevertheless, there also exist different possibilities for further improving particular processing stages of the WPDAM. In the presented implementation of WPDAM, the adaptive topology of the WPD tree using two automatically selectable empirically estimated basic topologies: one to analyze voiced speech and the other to analyze unvoiced speech. This approach could be, for example, further improved by using a precise phoneme classifier (not only rough voiced/unvoiced detection but also detailed classification of particular phoneme groups such as fricatives, stops, nasals, etc.) and to apply more specially designed wavelet packet tree topologies to precisely analyze those speech signal segments corresponding to these phoneme groups. Further improvements in automatic noise robust speech recognition performance can be achieved with improved multiresolutional signal analysis.

REFERENCES

[1] J.-C. Junqua and J. P. Haton, *Robustness in Automatic Speech Recognition*, Kluwer Academic, Boston, Mass, USA, 1996.

[2] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[3] M. J. F. Gales, *Model-based techniques for noise robust speech recognition*, Ph.D. thesis, University of Cambridge, Cambridge, UK, 1996.

[4] Y. Gong, "Speech recognition in noisy environments: a survey," *Speech Communication*, vol. 16, no. 3, pp. 261–291, 1995.

[5] ETSI standard document - ETSI ES 201 108 v1.1.1, "Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Front-end feature extraction algorithm, Compression algorithm," 2000.

[6] ETSI standard document - ETSI ES 202 050 v1.1.1, "Speech Processing, Transmission and Quality aspects (STQ), Distributed speech recognition, Advanced front-end feature extraction algorithm, Compression algorithm," 2002.

[7] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[8] H. Bourlard and S. Dupont, "Subband-based speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '97)*, vol. 2, pp. 1251–1254, Munich, Germany, April 1997.

[9] J. N. Gowdy and Z. Tufekci, "Mel-scaled discrete wavelet coefficients for speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, pp. 1351–1354, Istanbul, Turkey, June 2000.

[10] M. Gupta and A. Gilbert, "Robust speech recognition using wavelet coefficient features," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '01)*, pp. 445–448, Madonna di Campiglio, Trento, Italy, December 2001.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[12] K. K. Paliwal, "On the use of line spectral frequency parameters for speech recognition," *Digital Signal Processing*, vol. 2, no. 2, pp. 80–87, 1992.

[13] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, NY, USA, 1993.

[14] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Upper Saddle River, NJ, USA, 1993, section 4.5.

[15] R. R. Coifman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, part 2, pp. 713–718, 1992.

[16] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, Pa, USA, 1997.

[17] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674–693, 1989.

[18] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, Mass, USA, 1997.

[19] C.-T. Lu and H.-C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, no. 2-3, pp. 409–427, 2003.

[20] R. Sarikaya, B. L. Pellom, and J. H. L. Hansen, "Wavelet packet transform features with application to speaker identification," in *Proceedings of the 3rd IEEE Nordic Signal Processing Symposium (NORSIG '98)*, pp. 81–84, Vigsø, Denmark, June 1998.

[21] H. Sheikhzadeh and H. R. Abutalebi, "An improved wavelet-based speech enhancement system," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 1855–1858, Aalborg, Denmark, September 2001.

[22] K. Ramchandran, M. Vetterli, and C. Herley, "Wavelets, subband coding, and best bases," *Proceedings of the IEEE*, vol. 84, no. 4, pp. 541–560, 1996.

[23] N. R. Reyes, M. R. Zurera, F. L. Ferreras, and P. J. Amores, "Adaptive wavelet-packet analysis for audio coding purposes," *Signal Processing*, vol. 83, no. 5, pp. 919–929, 2003.

[24] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Processing Letters*, vol. 8, no. 1, pp. 10–12, 2001.

[25] D. L. Donoho, “De-noising by soft-thresholding,” *IEEE Transactions on Information Theory*, vol. 41, no. 3, pp. 613–627, 1995.

[26] E. Jafer and A. E. Mahdi, “Wavelet-based perceptual speech enhancement using adaptive threshold estimation,” in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH ’03)*, pp. 569–572, Geneva, Switzerland, September 2003.

[27] M. Jansen, *Wavelet thresholding and noise reduction*, Ph.D. thesis, Katholieke Universiteit Leuven, Leuven, Belgium, 2000.

[28] B. Andrassy, D. Vlaj, and C. Beaugeant, “Recognition performance of the siemens front-end with and without frame dropping on the aurora 2 database,” in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH ’01)*, pp. 193–196, Aalborg, Denmark, September 2001.

[29] H.-G Hirsch and D. Pearce, “The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in *Proceedings of the Automatic Speech Recognition: Challanges for the New Millennium (ISCA ITRW ASR ’00)*, pp. 181–188, Paris, France, September 2000.

[30] D. Pearce, “Enabling new speech driven services for mobile devices: an overview of the ETSI standards activities for distributed speech recognition front-ends,” in *Proceedings of Applied Voice Input/Output Society Conference (AVIOS ’00)*, San Jose, Calif, USA, May 2000.

[31] AU/225/00, “Baseline Results for Subset of SpeechDat-Car Finnish Database for ETSI STQ WI008 Advanced Front-end Evaluation,” Nokia, Janurary 2000.

[32] AU/271/00, “Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation: Description and Baseline Results,” UPC, November 2000.

[33] AU/273/00, “Description and Baseline Results for the Subset of the Speechdat-Car German Database used for ETSI STQ Aurora WI008 Advanced DSR Front-end Evaluation,” Texas Instruments, December 2001.

[34] AU/378/01, “Danish SpeechDat-Car Digits Database for ETSI STQ-Aurora Advanced DSR,” Aalborg University, January 2001.

[35] D. Macho, L. Mauuary, B. Noe, et al., “Evaluation of a noise-robust DSR front-end on aurora database,” in *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP ’02)*, pp. 17–20, Denver, Colo, USA, September 2002.

[36] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[37] B. Kotnik, D. Vlaj, and B. Horvat, “Efficient noise robust feature extraction algorithms for distributed speech recognition (DSR) systems,” *International Journal of Speech Technology*, vol. 6, no. 3, pp. 205–219, 2003.

[38] R. Martin, “Spectral subtraction based on minimum statistics,” in *Proceedings of the European Signal Processing Conference (EUSIPCO ’94)*, pp. 1182–1185, Edinburgh, UK, September 1994.

[39] D. O’Shaughnessy, “Enhancing speech degraded by additive noise or interfering speakers,” *IEEE Communications Magazine*, vol. 27, no. 2, pp. 46–52, 1989.

[40] J. H. McClellan and T. W. Parks, “A unified approach to the design of optimum FIR linear-phase digital filters,” *IEEE Transactions on Circuits Theory*, vol. 20, no. 6, pp. 697–701, 1973.

[41] O. Rioul and P. Duhamel, “A Remez exchange algorithm for orthonormal wavelets,” *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, vol. 41, no. 8, pp. 550–560, 1994.

[42] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[43] J. M. Hillenbrand and R. T. Gayvert, “Vowel classification based on fundamental frequency and formant frequencies,” *Journal of Speech and Hearing Research*, vol. 36, no. 4, pp. 694–700, 1993.

[44] M. Klein, “A Study of Voice Activity Detectors,” Speech Communications 304-523B, McGill University, Computer and Electrical Engineering Department, April 2000.

[45] B. Mak, J.-C. Junqua, and B. Reaves, “A robust speech/non-speech detection algorithm using time and frequency-based features,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’92)*, vol. 1, pp. 269–272, San Francisco, Calif, USA, March 1992.

[46] E. Nemer, R. Gourbran, and S. Mahmoud, “Robust voice activity detection using higher-order statistics in the LPC residual domain,” *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[47] J. Sohn and W. Sung, “A voice activity detector employing soft decision based noise spectrum adaptation,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’98)*, vol. 1, pp. 365–368, Seattle, Wash, USA, May 1998.

[48] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[49] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book—Version 3.0*, Microsoft, Redmond, Wash, USA, 2000.

[50] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *The Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[51] F. de Wet, B. Cranen, J. de Veth, and L. Boves, “A comparison of LPC and FFT-based acoustic features for noise robust ASR,” in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH ’01)*, pp. 865–868, Aalborg, Denmark, September 2001.

[52] R. Sarikaya and J. H. L. Hansen, “Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition,” in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH ’01)*, pp. 687–690, Aalborg, Denmark, September 2001.

[53] B. Kotnik, Z. Kačič, and B. Horvat, “Development and integration of the LDA-toolkit into the COST249 speechdat (II) SIG reference recognizer,” in *Proceedings the 4th International Conference on Language Resources and Evaluation (LREC ’04)*, pp. 2083–2086, Lisbon, Portugal, May 2004.

[54] L. Welling, *Merkmalsextraction in spracherkennungssystemen für grossen wortschatz*, Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule, Aachen, Germany, 1999.

Bojan Kotnik received the Diploma degree from University of Maribor in 2000 and the Ph.D. degree in 2004 at the same institution. Since 2004 he has been a Senior Researcher at Faculty of Electrical Engineering and Computer Science at the University of Maribor. His research interests are in the areas of feature extraction, feature postprocessing, speech enhancement, pitch determination, and robust speech recognition.

Zdravko Kačič graduated at the Faculty of Electrical Engineering and Computer Science at University of Maribor in 1986. He was awarded the M.S. degree in 1989 and the Ph.D. degree in 1992 at the same faculty, where he is currently employed as Full Professor. He is the head of the Laboratory for Digital Signal Processing. His research interests are the analysis of the complex sound scenes, systems for automatic speech recognition, and the creation of language resources.